策略梯度一个公式的推导

所有轨迹下，t 时刻的 reward r_t 的期望 $E_\tau r_t$ 的策略梯度

$$\nabla_\theta E_\tau r_t = \nabla_\theta \sum_\tau p_\theta(\tau) r_t =$$

$$\sum_\tau \nabla_\theta p_\theta(\tau) r_t = \sum_\tau p_\theta(\tau) \nabla_\theta log\{p_\theta(\tau)\} r_t =$$

$$\sum_\tau p_\theta(\tau) \{\sum_{i=0}^{T-1} \nabla_\theta log\{\pi_\theta(a_i|t_i)\}\} r_t$$

对中间的梯度之和，只取第 $t+k$ 项, $k>0$

$$\sum_\tau p_\theta(\tau) \nabla_\theta log\{\pi_\theta(a_{t+k}|s_{t+k})\} r_t =$$

$$\sum_{s_0, a_0 .. s_{T-1}, a_{T-1}} p(s_0) \cdot \prod_{i=0}^{T-1} \pi_\theta(a_i|s_i) \cdot p(s_{i+1}|s_i, a_i) \nabla_\theta log\{\pi_\theta(a_{t+k}|s_{t+k})\} r_t =$$

$$\sum_{s_0, a_0 .. s_{t+k}, a_{t+k}} \sum_{s_{t+k+1}, a_{t+k+1} .. s_{T-1}, a_{T-1}} p(s_0) \cdot \prod_{i=0}^{T-1} \pi_\theta(a_i|s_i) \cdot p(s_{i+1}|s_i, a_i) \cdot$$

$$\nabla_\theta log\{\pi_\theta(a_{t+k}|s_{t+k})\} r_t =$$

$$\sum_{s_0, a_0 .. s_{t+k}, a_{t+k}} p(s_0) \cdot \prod_{i=0}^{t+k-1} \pi_\theta(a_i|s_i) \cdot p(s_{i+1}|s_i, a_i) \cdot \pi_\theta(a_{t+k}|s_{t+k}) \nabla_\theta log\{\pi_\theta(a_{t+k}|s_{t+k})\} r_t \cdot$$

$$\sum_{s_{t+k+1}, a_{t+k+1} .. s_{T-1}, a_{T-1}} p(s_{i+1}|s_i, a_i) \cdot \prod_{i=t+k+1}^{T-1} \pi_\theta(a_i|s_i) \cdot p(s_{i+1}|s_i, a_i)$$

考察

$$\sum_{s_0, a_0 .. s_{t+k}, a_{t+k}} p(s_0) \cdot \prod_{i=0}^{t+k-1} \pi_\theta(a_i|s_i) \cdot p(s_{i+1}|s_i, a_i) \cdot \pi_\theta(a_{t+k}|s_{t+k}) \nabla_\theta log\{\pi_\theta(a_{t+k}|s_{t+k})\} r_t =$$

$$\sum_{s_0, a_0 .. s_{t+k}} r_t \cdot p(s_0) \cdot \prod_{i=0}^{t+k-1} \pi_\theta(a_i|s_i) \cdot p(s_{i+1}|s_i, a_i) \cdot \sum_{a_{t+k}} \pi_\theta(a_{t+k}|s_{t+k}) \nabla_\theta log\{\pi_\theta(a_{t+k}|s_{t+k})\}$$

给定s_{t+k},其中

$$\sum_{a_{t+k}} \pi_\theta(a_{t+k}|s_{t+k}) \nabla_\theta log\{\pi_\theta(a_{t+k}|s_{t+k})\} =$$

$$\sum_{a_{t+k}} \pi_\theta(a_{t+k}|s_{t+k}) \frac{\nabla_\theta \pi_\theta(a_{t+k}|s_{t+k})}{\pi_\theta(a_{t+k}|s_{t+k})} = \sum_{a_{t+k}} \nabla_\theta \pi_\theta(a_{t+k}|s_{t+k}) =$$

$$\nabla_\theta \sum_{a_{t+k}} \pi_\theta(a_{t+k}|s_{t+k}) = \nabla_\theta 1 = 0$$

从而可知，当 $k > 0$

$$\sum_\tau p_\theta(\tau) \nabla_\theta log\{\pi_\theta(a_{t+k}|s_{t+k})\} r_t = 0 \Longrightarrow$$

$$\sum_\tau p_\theta(\tau)\{\sum_{i=0}^{T-1} \nabla_\theta log\{\pi_\theta(a_i|t_i)\}\} r_t = \sum_\tau p_\theta(\tau)\{\sum_{i=0}^{t} \nabla_\theta log\{\pi_\theta(a_i|t_i)\}\} r_t \Longrightarrow$$

$$\nabla_\theta E_\tau r_t = \sum_\tau p_\theta(\tau)\{\sum_{i=0}^{t} \nabla_\theta log\{\pi_\theta(a_i|t_i)\}\} r_t$$

即 t 时刻的奖励的期望的策略梯度与 t 时刻之后的各时刻的策略梯度无关。